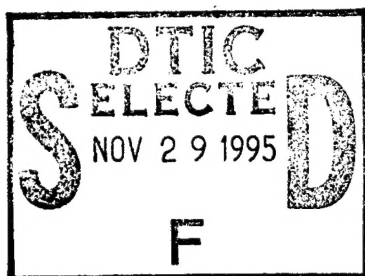


NPS-OR-95-009

NAVAL POSTGRADUATE SCHOOL Monterey, California



RANDOMIZED QUANTILE RESIDUALS

by

Peter K. Dunn
Gordon K. Smyth

September 1995

Approved for public release; distribution is unlimited.

Prepared for: Naval Postgraduate School
Monterey, CA

DTIC QUALITY INSPECTED 5

19951127 068

19951127 068

**THERE ARE NO PAGES MISSING, PAGE 10
IS THE LAST PAGE OF THE REPORT. PAGE 15
WAS MISNUMBERED, IT SHOULD BE PAGE 11.
PAGES 12, 13 & 14 DONOT EXIST PER FRANCES
SMITH (DSN: 878-2061) NAVAL POSTGRADUATE
SCHOOL, MONTEREY, CA.
JANUARY 22, 1996**

NAVAL POSTGRADUATE SCHOOL
MONTEREY, CA 93943-5000

Rear Admiral M. J. Evans
Superintendent


Richard Elster
Provost

This report was prepared for and funded by the Naval Postgraduate School.

Reproduction of all or part of this report is authorized.

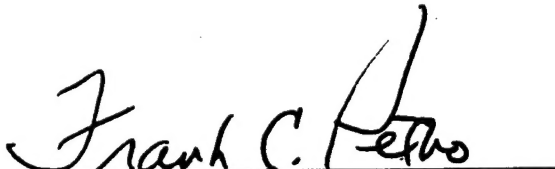
This report was prepared by:


PETER K. DUNN
Ph.D. Student


GORDON K. SMYTH
Professor of Operations Research

Reviewed by:

Released by:


FRANK PETHO
Acting Chairman
Department of Operations Research


PAUL J. MARTO
Dean of Research

Accession For		
NTIS	CRA&I	<input checked="checked" type="checkbox"/>
DTIC	TAB	<input type="checkbox"/>
Unannounced		<input type="checkbox"/>
Justification		
By		
Distribution /		
Availability Codes		
Dist	Avail and/or Special	
A-1		

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.				
1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE 28 Sep 95		3. REPORT TYPE AND DATES COVERED Technical
4. TITLE AND SUBTITLE Randomized Quantile Residuals			5. FUNDING NUMBERS	
6. AUTHOR(S) Peter K. Dunn and Gordon K. Smyth				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943			8. PERFORMING ORGANIZATION REPORT NUMBER NPS-OR-95-009	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943			10. SPONSORING / MONITORING AGENCY REPORT NUMBER	
11. SUPPLEMENTARY NOTES				
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.			12b. DISTRIBUTION CODE	
13. ABSTRACT (Maximum 200 words) In this paper we give a general definition of residuals for regression models with independent responses. Our definition produces residuals which are exactly normal, apart from sampling variability in the estimated parameters, by inverting the fitted distribution function for each response value and finding the equivalent standard normal quantile. Our definition includes some randomization to achieve continuous residuals when the response variable is discrete. Quantile residuals are easily computed in computer packages such as SAS, S-Plus, GLIM or LispStat, and allow residual analyses to be carried out in many commonly occurring situations in which the customary definitions of residuals fail. Quantile residuals are applied in this paper to three example data sets.				
14. SUBJECT TERMS generalized linear model; deviance residual; Pearson residual; exponential regression; logistic regression; Poisson regression; normal probability plot.			15. NUMBER OF PAGES 14	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT UL	

Randomized Quantile Residuals

Peter K. Dunn and Gordon K. Smyth*

Department of Mathematics, University of Queensland,
Brisbane, Q 4072, Australia.

September 28, 1995

Abstract

In this paper we give a general definition of residuals for regression models with independent responses. Our definition produces residuals which are exactly normal, apart from sampling variability in the estimated parameters, by inverting the fitted distribution function for each response value and finding the equivalent standard normal quantile. Our definition includes some randomization to achieve continuous residuals when the response variable is discrete. Quantile residuals are easily computed in computer packages such as SAS, S-Plus, GLIM or LispStat, and allow residual analyses to be carried out in many commonly occurring situations in which the customary definitions of residuals fail. Quantile residuals are applied in this paper to three example data sets.

Keywords: generalized linear model; deviance residual; Pearson residual; exponential regression; logistic regression; Poisson regression; normal probability plot.

1 Introduction

Residuals, and especially plots of residuals, play a central role in the checking of statistical models. In normal linear regression the residuals are normally distributed and can be standardized to have equal variances. In non-normal regression situations, such as logistic regression or log-linear analysis, the residuals, as usually defined, may be so far from normality and from having equal variances as to be of no practical use. A particular problem occurs when the response variable is discrete and takes on a small number of

*Current address Department of Operations Research, Naval Postgraduate School, Monterey, CA 93943

distinct values, as for Poisson data with mean not far from zero or binomial data with mean close to either zero or the number of trials. In such situations the residuals lie on parallel curves corresponding to distinct response values, and these spurious curves distract the eye seriously from any meaningful message that might be contained in a residual plot.

In this paper we give a general definition of residuals for regression models with independent responses. Our definition produces residuals which are exactly normal, apart from sampling variability in the estimated parameters, by inverting the fitted distribution function for each response value and finding the equivalent standard normal quantile. This approach is closely related to that of Cox and Snell (1968), but whereas Cox and Snell concentrate on mean and variance corrections we concentrate on the transformation to normality. Our definition includes some randomization to achieve continuous residuals when the response variable is discrete. Quantile residuals are easily computed in computer packages such as SAS, S-Plus, GLIM or LispStat, and allow residual analyses to be carried out in many commonly occurring situations in which the customary definitions of residuals fail.

For other work on residuals for non-normal regression models see Pierce and Schafer (1986) or McCullagh and Nelder (1989) and the references therein. In the discussion at the end of the paper we briefly indicate how quantile residuals may be extended to models with dependent responses.

2 Pearson and Deviance Residuals

Let y_1, \dots, y_n be responses and for each i let \mathbf{x}_i be a vector of covariates. The y_i are assumed to be independent and to follow a distribution $\mathcal{P}(\mu_i, \phi)$ where $\mu_i = E(y_i)$ and ϕ is a parameter vector common to all the y_i . The μ_i are assumed to depend on the \mathbf{x}_i and a vector of regression parameters β . We have particularly in mind generalized linear

models (McCullagh and Nelder, 1989) in which the probability density or mass function of y_i has the form

$$f(y; \theta_i, \phi) = a(y, \phi) \exp[\{y\theta_i - \kappa(\theta_i)\}/\phi]$$

where $a()$ and $\kappa()$ are known functions and $\mu_i = \kappa'(\theta_i)$. In this model we have $\text{var}(y_i) = \phi V(\mu_i)$ where $V(\mu_i) = \kappa''(\theta_i)$. It is customary to assume that $g(\mu_i) = \mathbf{x}^T \boldsymbol{\beta}$ where $g()$ is a known link function. The parameter ϕ is the proportionality constant in the mean-variance relationship and is known as the dispersion parameter.

In the context of generalized linear models, two definitions of residuals have been commonly used in practice. The Pearson residual is defined by

$$r_{p,i} = \frac{y_i - \hat{\mu}_i}{V(\hat{\mu}_i)^{1/2}}$$

where $\hat{\mu}_i$ is the fitted value for μ_i . The Pearson residual has the advantage that its mean and variance are exactly zero and ϕ respectively, if sampling variability in $\hat{\mu}_i$ is small. The deviance residuals are defined in terms of the unit deviances. For the above model, let $t(y, \mu) = y\theta - \kappa(\theta)$. Assuming that y is in the domain of μ , the unit deviance is

$$d(y, \mu) = 2\{t(y, y) - t(y, \mu)\}$$

The deviance residual is

$$r_{d,i} = d(y_i, \hat{\mu}_i)^{1/2} \text{sign}(y_i - \hat{\mu}_i)$$

Pierce and Schafer (1986) have argued on theoretical grounds that the deviance residuals should be more nearly normal than the Pearson. Indeed both converge to normality as $\phi \rightarrow 0$ relative to the μ_i , the Pearson residuals at rate $O(\phi^{1/2})$ by the Central Limit Theorem and the deviance residuals at $O(\phi)$ by the saddle-point approximation to $f(y; \theta_i, \phi)$. The Pearson and deviance residuals coincide and are exactly normal, ignoring variability in $\hat{\mu}_i$, for the normal linear model. The deviance residual is also exactly normal when the response is inverse-Gaussian. In other cases and for large ϕ/μ however, neither type

of residual can be guaranteed to be closely normal, and the deviance residuals do not generally have zero means or equal variances even at the true values μ_i .

3 Randomized Quantile Residuals

Let $F(y; \mu_i, \phi)$ be the cumulative distribution function of $\mathcal{P}(\mu_i, \phi)$. If F is continuous, the quantile residuals are defined by

$$r_{q,i} = \Phi^{-1}\{F(y_i; \hat{\mu}_i, \hat{\phi})\}$$

where $\Phi()$ is the cumulative distribution function of the standard normal. Apart sampling variability in $\hat{\mu}_i$ and $\hat{\phi}$, the $r_{q,i}$ are exactly standard normal. This implies that the distribution of $r_{q,i}$ converges to standard normal if β and ϕ are consistently estimated. The above definition is a special case of Cox and Snell's (1968) "crude" residuals.

Example 1: Leukemia data. Feigl and Zelen (1965) discuss some data relating the survival times y_i of leukemia patients to their initial white blood cell counts x_i and to existence of AG-factor. Following Feigl and Zelen, we treat the survival times as exponential, $y_i \sim \text{Exp}(\mu_i)$. We work with a log-linear model for the means, including separate intercepts for the two AG-factor groups,

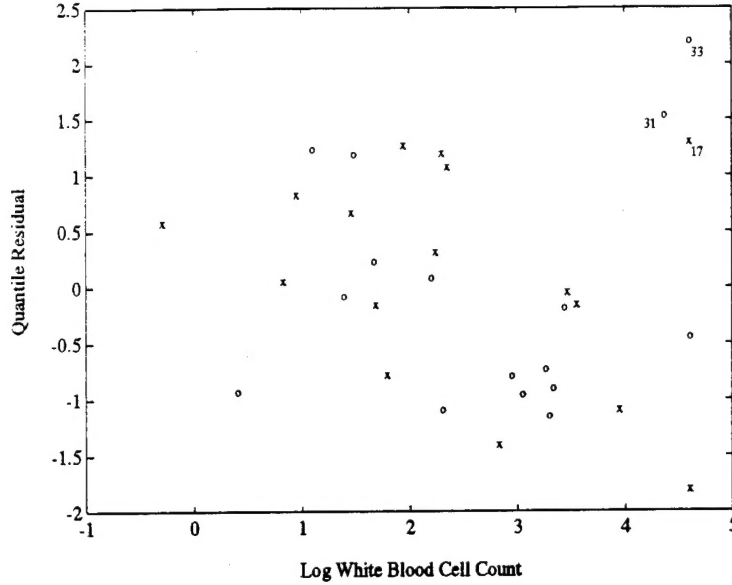
$$\log \mu_i = \begin{cases} \alpha_1 + \beta \log x_i & \text{AG positive} \\ \alpha_2 + \beta \log x_i & \text{AG negative} \end{cases}$$

Cox and Snell (1968) considered a subset of this data, and defined approximately exponential crude residuals $R_i = y_i/\hat{\mu}_i$, where the $\hat{\mu}_i$ are the estimated means. In this case the quantile residuals

$$r_{q,i} = \Phi^{-1}\{1 - \exp(y_i/\hat{\mu}_i)\}$$

are a simple transformation of the R_i . A normal probability plot of the quantile residuals confirms the assumption of an exponential distribution. Figure 1 plots the quantile residuals versus the covariate. The three residuals (cases 17, 31 and 33) in the upper

Figure 1: Plot of quantile residuals versus the covariate for the leukemia data. Circles represent patients which are AG-positive, crosses AG-negative.



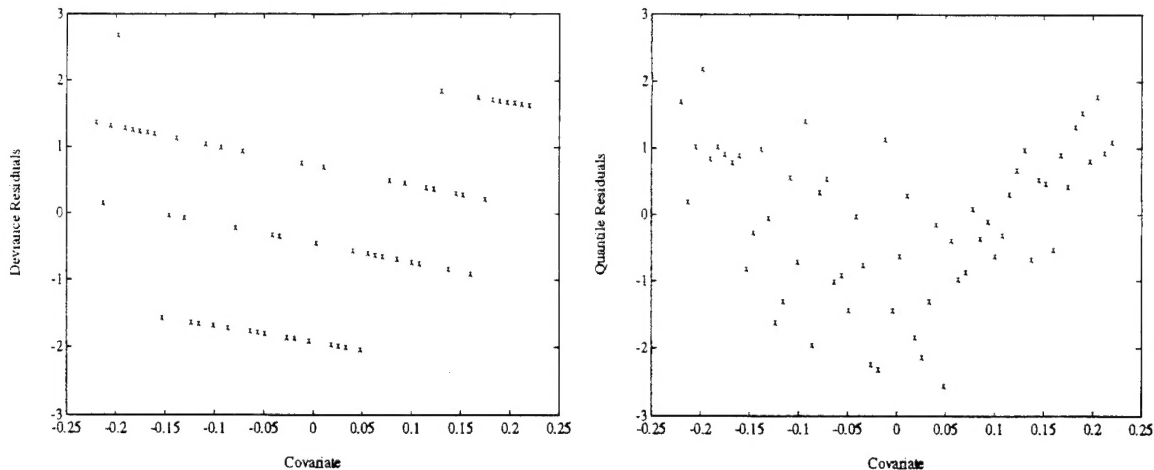
right-hand corner of the plot are relatively separate from the body of the other residuals, and without them there appears to be a marked negative trend. Cases 17, 31 and 33 may be outliers, or it may be that the dispersion of the residuals increases at the largest white blood cell counts. In any case, the three cases identified appear from the residual plot to be jointly influential. Assigning the identified cases zero weight increases $\hat{\beta}$ nearly three-fold, from -0.30 to -0.84 compared with a standard error of 0.14.

If F is not continuous, a more general definition of quantile residuals is required. Let $a_i = \lim_{y \uparrow y_i} F(y; \hat{\mu}_i, \hat{\phi})$ and $b_i = F(y_i; \hat{\mu}_i, \hat{\phi})$. We define the randomized quantile residual for y_i by

$$r_{q,i} = \Phi^{-1}(u_i)$$

where u_i is a uniform random variable on the interval $(a_i, b_i]$. Again, the $r_{q,i}$ are exactly standard normal, apart sampling variability in $\hat{\mu}_i$ and $\hat{\phi}$. The randomization strategy employed here is similar to the strategy of jittering (Chambers et al, 1983) to prevent masses of overlapping points in plots. Whereas jittering applies a uniform random component to

Figure 2: Deviance and quantile residuals versus the covariate from a logistic regression. The response is simulated $\text{bin}(3, p)$ with $\text{logit } p$ depending quadratically on the covariate.

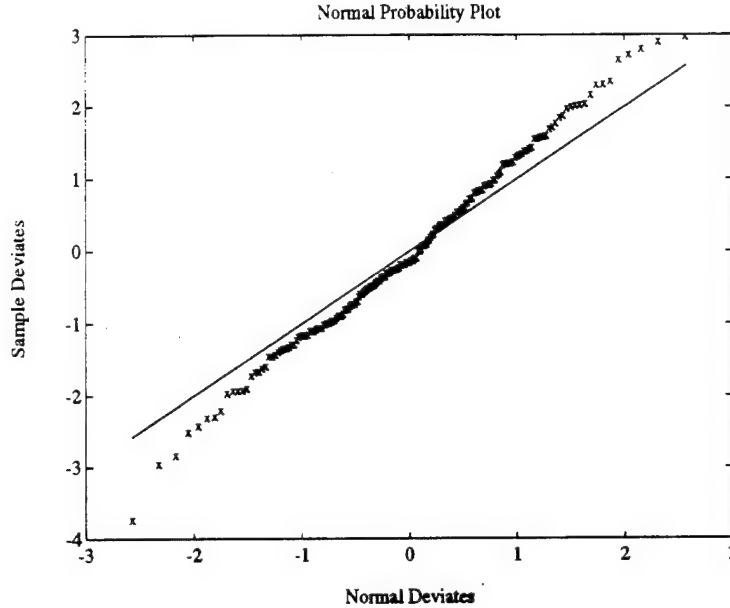


the response, our uniform random component is on the cumulative probability scale and is tailored to the actual probability mass at the point in question. Our randomization is the minimum necessary so that no granularity remains in the resulting residual distribution.

Example 2: Simulated binomial data. A logistic linear regression was used to model 60 binomial observations with binomial denominator $n = 3$, i.e., the responses were assumed to be independently distributed as $y_i \sim \text{bin}(n, p_i)$, with $n = 3$ and $\text{logit}(p_i) = \beta_0 + \beta_1 x_i$ where x_i is a covariate. The first plot of Figure 2 displays the deviance residuals versus the covariate. The points in this plot lie on four parallel curves corresponding to the four possible values for the response. The curves make it difficult to see any other pattern in the data. The second plot displays the quantile residuals versus the covariate. In this plot it is clear that the residuals follow a quadratic pattern. The data for this example was in fact computer generated with $\text{logit}(p_i)$ depending quadratically on the x_i .

Example 3: Fathers' and sons' occupations. Brown (1974) and Kotze and Hawkins (1984) analyze a sparse 14×14 contingency table showing the cross-classification of occupations of fathers (rows) by occupations of sons (columns). The data was originally published by Pearson (1904) and appears also in Hand et al (1994). Brown, Kotze and

Figure 3: Normal probability plot of the quantile residuals from the fathers' and sons' occupation data.



Hawkins were interested in identifying those cells which are outliers relative to the independence model. We take a similar approach, with the difference that the quantile residual approach allows us to look for outliers relative to a more realistic model. Observing that there is an apriori expectation that sons will be influenced by their father's occupation, we fit a log-linear Poisson regression model to the counts with row and column effects and with an effect for equality of father's and son's occupation, i.e., $y_{ij} \sim \text{Pois}(\mu_{ij})$, with

$$\log \mu_{ij} = \mu_0 + \alpha_i + \beta_j + \delta x_{ij} \quad (1)$$

and $x_{ij} = 1$ if $i = j$ and 0 otherwise. Figure 3 is a normal probability plot of quantile residuals from this model. The largest positive residual corresponds to the (2,2) cell: sons almost always continue to work in the Arts if their father did. Figure 3 shows evidence of large negative residuals as well as large positive residuals. Although none of the negative residuals are individually significant, and the actual contingency table cells represented in the left tail of the probability plot varies with each realization of the quantile residuals, the overall pattern is preserved across realizations. The quantile residual plot shows in this

way that there are too many zero counts in the contingency table to be compatible with the above model. No other method which has been applied to this data in the literature is able to show this aspect of the data. Although Figure 3 shows clear evidence of lack of fit, the model (1) and the models which arise from it by deleting selected cells does give an appreciably better fit to this data than the independence models considered by earlier authors.

4 Discussion and Extensions

In this paper quantile residuals are computed by finding the equivalent standard normal deviate for each response observation. Any reference distribution could have been chosen for the residuals, but the normal seems to be the easiest to interpret for graphical purposes.

Randomization is used to produce continuously distributed residuals when the response is discrete or has a discrete component. This means that the quantile residuals will vary from one realization to another for a given data set and fitted model. For the sake of brevity, we have given only one realization of the quantile residuals for each example in this paper. In practice though we have found it useful to routinely plot four realizations of the quantile residuals. Any pattern in the residuals which is not consistent across the realizations is then ignored.

Quantile residuals provide a logical approach to added variable plots (Cook and Weisberg, 1982) in generalized linear models. An added variable plot for a variable x would consist of plotting the quantile residuals, for the model excluding x , versus x_a , where x_a is x adjusted for the other covariates in the model. The vector x_a would be chosen to be orthogonal to the other covariates, relative to the covariance matrix of the y_i . It might be computed as the residuals from weighted least squares regression of x on the other covariates, using as weights the working weights from the generalized linear model.

Independence of the response observations was assumed in this paper. The method

of quantile residuals can be extended to dependent data situations by expressing the multivariate likelihood as a sum of univariate conditional likelihoods. For example we might define the i th conditional quantile residual from the conditional distribution of y_i given y_1, \dots, y_n instead of from the marginal distribution of y_i as in the paper. This would provide independent, standard normal residuals.

Finally we consider the sampling variability of the $\hat{\mu}_i$, which has for simplicity been ignored throughout this paper. Treating the $\hat{\mu}_i$ as fixed is appropriate when good information is available on the model parameters, but may be unrealistic for example for designed experiments in which the number of parameters is not small compared to the number of observations. In normal linear models, REML estimate of the variance structure is obtained from the marginal distribution of any set of zero mean contrasts, $Z^T y$ say. In a similar way, independent and identically distributed residuals could be obtained by transforming from the y_i to any orthonormal set of zero mean contrasts.

Extending this idea to non-normal regression is more difficult, but could in principle be done using the conditional approach of Smyth and Verbyla (1995). Smyth and Verbyla (1995) have argued that REML estimation for generalized linear models should proceed by considering the conditional distribution of the y_i given $\hat{\beta}$. Independent quantile residuals could therefore be defined by considering the conditional distribution of each y_i given y_1, \dots, y_{i-1} and $\hat{\beta}$. For certain values of i this distribution would be degenerate; these values could be ignored without loss of information.

References

- Brown, M. B. (1974). Identification of the sources of significance in two-way contingency tables. *Appl. Statist.*, **23**, 405-413.
- Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983). *Graphical Methods for Data Analysis*, Wadsworth, Belmont, California.

- Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman and Hall, New York.
- Cox, D. R. and Snell, E. J. (1968). A general definition of residuals (with discussion). *J. R. Statist. Soc., B*, **30**, 248–275.
- Feigl, P. and Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant observation. *Biometrics*, **21**, 826–838.
- Hand, D. J. *et al* (1994). *Handbook of Small Data Sets*. Chapman & Hall, London.
- Kotze, T. J. v W. and Hawkins, D. M. (1984). The identification of outliers in two-way contingency tables using 2×2 subtables. *Appl. Statist.*, **33**, 215–223.
- McCullagh, P. and Nelder, J.A. (1989). *Generalized linear models, 2nd ed.* Chapman and Hall: London.
- Pearson, K. (1904). On the theory of contingency and its relation to association and normal correlation. Reprinted in 1948 in *Karl Pearson's Early Statistical Papers*, Cambridge University Press, Cambridge, 443–475.
- Pierce, D. A., and Schafer, D. W. (1986). Residuals in generalized linear models. *J. Amer. Statist. Ass.*, **81**, 977–986.
- Pregibon, D. (1981). Logistic regression diagnostics. *Ann. Statist.*, **9**, 705–724.
- Smyth, G. K. and Verbyla, A. P. (1995). A conditional approach to REML in generalized linear models. *J. Roy. Statist. Soc. B*. To appear, 12 pages.

DISTRIBUTION LIST

1. Research Office (Code 08) 1
Naval Postgraduate School
Monterey, CA 93943-5000
2. Dudley Knox Library (Code 52) 2
Naval Postgraduate School
Monterey, CA 93943-5002
3. Defense Technical Information Center 2
Cameron Station
Alexandria, VA 22314
4. Department of Operations Research 1
Editorial Assistant (Code OR/Bi)
Naval Postgraduate School
Monterey, CA 93943-5000
5. Prof. Gordon K. Smyth (Code OR/Sg) 5
Naval Postgraduate School
Monterey, CA 93943-5000